The Weak Aggregating Algorithm and Weak Mixability ⋆

Yuri Kalnishkan and Michael V. Vyugin

Department of Computer Science and Computer Learning Research Centre, Royal Holloway, University of London, Egham, Surrey, TW20 0EX, United Kingdom

Abstract

This paper resolves the problem of predicting as well as the best expert up to an additive term of the order $o(n)$, where n is the length of a sequence of letters from a finite alphabet. We call the games that permit this weakly mixable and give a geometrical characterisation of the class of weakly mixable games. Weak mixability turns out to be equivalent to convexity of the finite part of the set of superpredictions. For bounded games we introduce the Weak Aggregating Algorithm that allows us to obtain additive terms of the form $C\sqrt{n}$.

Key words: on-line learning, predicting individual sequences, prediction with expert advice, general loss functions
1991 MSC: 68T05, 68Q32

1 Introduction

This paper deals with the problem of prediction with expert advice. We consider the on-line prediction protocol, where outcomes $\omega_1, \omega_2, \ldots$ occur in succession while a prediction strategy tries to predict them. Before seeing an event ω_t, the prediction strategy produces a prediction γ_t. We are interested in the case of a finite outcome space, i.e., $\omega_1, \omega_2, \ldots \in \Omega$ such that $|\Omega| < +\infty$.

⋆ The previous versions of this paper were published as Technical Report CLRC-TR-03-01, Computer Learning Research Centre, Royal Holloway, University of London (November 2003) and in *Learning Theory, Proceedings of the 18th Annual Conference (COLT 2005)*, volume 3559 of *Lecture Notes in Artificial Intelligence*, Springer, 2005.

Email addresses: yura@cs.rhul.ac.uk (Yuri Kalnishkan), misha@cs.rhul.ac.uk (Michael V. Vyugin).

We use a loss function $\lambda(\omega, \gamma)$ to measure the discrepancies between predictions and outcomes. A loss function and a prediction space (a set of possible predictions) Γ specify the game, i.e., a particular prediction environment. The performance of a learner $\mathfrak{S}$ w.r.t. a game is measured by the cumulative loss

$$\mathrm{Loss}_{\mathfrak{S}}(n) = \sum_{t=1}^{n} \lambda(\omega_t, \gamma_t) \ . \tag{1}$$

suffered on a sequence of outcomes $\omega_1, \omega_2, \ldots, \omega_n$. In the problem of prediction with expert advice the learner has access to predictions generated by 'experts' $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_N$ that try to predict elements of the same sequence. The goal of the learner is to predict nearly as well as the best expert, i.e., to suffer loss that is only little bigger than the smallest of the experts' losses.

This problem has been studied intensively; see, e.g., [1,2] and the overview in the recently published book [3]. Papers [4,5] propose the Aggregating Algorithm that allows the learner $\mathfrak{M}$ to achieve loss satisfying the inequality

$$\mathrm{Loss}_{\mathfrak{M}}(n) \le c\,\mathrm{Loss}_{\mathcal{E}_i}(n) + a \ln N \tag{2}$$

for all $i = 1 \ldots, N$ and $n = 1, 2, \ldots$, where the constants c and a are optimal and are specified by the game. Note that neither c nor a depend on n.

If we can take c equal to 1, the game is called mixable. It is possible to provide a geometrical characterisation of mixable games in terms of the so called sets of superpredictions. The Aggregating Algorithm fully resolves the problem of predicting as well as the best expert up to an additive constant. For the sake of completeness we formulate one of the results concerning the Aggregating Algorithm in Subsection 2.4.

There are interesting games that are not mixable, e.g., the absolute loss game introduced in Subsection 2.1. The Aggregating Algorithm still works for some of such games, but it allows us to achieve only values of c greater than 1.

In this paper we take a different approach to non-mixable games. We fix $c = 1$ but consider $a(n)$ that can grow when the length n of the sequence increases. We study the problem of predicting as well as the best expert up to $o(n)$ as $n \to +\infty$, where n is the length of the sequence. Section 3 introduces the corresponding concept of weak mixability. The main result of this paper, Theorem 7, shows that weak mixability is equivalent to a very simple geometric property of the set of superpredictions, namely, the convexity of its finite part.

If the loss function is bounded, it is possible to predict as well as the best expert up to an additive term of the form $C\sqrt{n}$, provided the finite part of the set of superpredictions is convex. This result follows from a recent paper [6]. We shall present our own construction, which is independent of [6] and goes

back to ideas from [1]. We develop the Weak Aggregating Algorithm based on the old method of averaging experts' losses with dynamically updated weights. Unlike the Aggregating Algorithm, which uses the average $\log_\beta(\sum_i p_i \beta^{l_i})$, the Weak Aggregating Algorithm uses simple convex combinations $\sum_i p_i l_i$ (see Remark 19 for a more detailed discussion); however the way of updating weights is more complicated.

In [6] (see Remark 'Deterministic prediction and absolute loss' at the end of Section 9 in [6]) a result similar to Corollary 14 was obtained. The algorithm used in [6] is based on the 'following the perturbed leader' idea while ours belongs to the family of 'weighted majority'-type algorithms. The extra term obtained in [6], Theorem 6.*ii* has the multiplicative constant $2\sqrt{2}$ as compared to our 2. On the other hand the analysis in [6] is much more general; some of the bounds obtained there have extra terms depending on the loss of experts rather than time. Those bounds make sense even when the experts' loss is small, while ours is meaningful only for big losses.

Different algorithms and results leading to various extra terms are widely discussed in the literature; for an overview see [3], Chapter 2 including bibliographic remarks in 2.12. The general question of lower bounds for additive terms of the type considered in this paper remains open. The authors derive some lower bounds in [7] (that paper deals with predictive complexity, but the results can be easily restated for the problem of prediction with expert advice) but the bounds are not sufficiently tight.

If the game is not bounded, our construction can be applied in a different form to predict as well as the best expert up to $o(n)$. The result for unbounded games as well as the negative result for games that are not convex constitute the most original contribution of the paper (Appendix A shows that there are indeed unbounded games that are convex but not mixable).

The question of lower bounds for the additive term for unbounded games remains open too.

2 Preliminaries

We will formulate the definitions below without a reference to computability. The negative results of this paper are true in this strong sense. The positive results are proved constructively and algorithms are presented. Therefore the theory can be reformulated in a constructive fashion; see Section 7 for details.

A *game* $\mathfrak{G}$ is a triple $\langle \Omega, \Gamma, \lambda \rangle$, where Ω is an *outcome space*, Γ is a *prediction space*, and $\lambda : \Omega \times \Gamma \to [0, +\infty]$ is a *loss function*. We assume that Ω is a finite set of cardinality $M < +\infty$; we shall refer to elements of Ω as to $\omega^{(0)}, \omega^{(1)}, \ldots, \omega^{(M-1)}$. In the simplest binary case $M = 2$ and Ω may be identified with $\mathbb{B} = \{0, 1\}$. We also assume that Γ is a compact topological space and λ is continuous w.r.t. the extended topology of $[-\infty, +\infty]$. Since we treat Ω as a discrete space, the continuity of λ in two arguments is the same as continuity in the second argument. These assumption hold throughout the paper except for Remark 8, where negative losses are discussed.

The *square-loss* game, the *absolute-loss* game, and the *logarithmic* game with the outcome space $\Omega = \mathbb{B}$, prediction space $\Gamma = [0, 1]$, and loss functions $\lambda(\omega, \gamma) = (\omega - \gamma)^2$, $\lambda(\omega, \gamma) = |\omega - \gamma|$, and

$$\lambda(\omega, \gamma) = \begin{cases} -\log_2(1 - \gamma) & \text{if } \omega = 0 \ , \\ -\log_2 \gamma & \text{if } \omega = 1 \ , \end{cases}$$

respectively, are examples of (binary) games. A slightly different example is provided by the *simple prediction game* with $\Omega = \Gamma = \mathbb{B} = \{0, 1\}$ and $\lambda(\omega, \gamma) = 0$ if $\omega = \gamma$ and $\lambda(\omega, \gamma) = 1$ otherwise.

A game $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ is *bounded* if and only if λ is bounded, i.e., there is $L \in (0, +\infty)$ such that $\lambda(\omega, \gamma) \le L$ for each $\omega \in \Omega$ and $\gamma \in \Gamma$. If a game is not bounded, we shall call it *unbounded*. Examples of bounded games include the square-loss game, the absolute-loss game, and the simple prediction game. The logarithmic game is unbounded.

It is essential to allow λ to assume the value $+\infty$; this assumption is necessary in order to take into account the logarithmic game as well as other unbounded games. However we impose the following restriction: if $\lambda(\omega_0, \gamma_0) = +\infty$ for some $\omega_0 \in \Omega$ and $\gamma_0 \in \Gamma$, then there is a sequence $\gamma_n \in \Gamma$ such that $\gamma_n \to \gamma_0$ and $\lambda(\omega, \gamma_n)$ is finite for all $\omega \in \Omega$ and all positive integers n (note that $\lambda(\omega_0, \gamma_n) \to +\infty$ by continuity). In other words, any prediction that leads to infinite loss on some outcomes can be approximated by predictions that can only lead to finite loss no matter what outcome occurs. This restriction allows us to exclude some degenerate cases and to simplify the statements of theorems.

A merging strategy works in an on-line fashion. On trial t it reads predictions of experts $\mathcal{E}^{(1)}, \mathcal{E}^{(2)}, \ldots, \mathcal{E}^{(N)}$ and outputs its own. After ω_t, the outcome of trial t, becomes available, the experts and the merging strategy suffer losses.

We want the merging strategy to compete with the experts in terms of the cumulative loss. The goal of the merging strategy is to suffer loss that is not much worse than the loss of the best expert. By the best expert after trial t we mean the expert that has suffered the smallest cumulative loss so far.

Formally a *merging strategy* $\mathfrak{M}$ *for* N *experts* is a function

$$\mathfrak{M} : \bigcup_{t=1}^{+\infty} \left(\Omega^{t-1} \times \left(\Gamma^N \right)^t \right) \to \Gamma \ . \tag{3}$$

Consider the following on-line protocol:

```
(1) FOR t = 1, 2, ...
(2)     𝔐 reads predictions γ_t^(1), γ_t^(2), ..., γ_t^(N) ∈ Γ
(3)     𝔐 chooses γ_t ∈ Γ
(4)     𝔐 observes the actual outcome ω_t ∈ Ω
(5) END FOR
```

By definition, let the total loss of $\mathfrak{M}$ after n trials be

$$\mathrm{Loss}_{\mathfrak{M}}^{\mathfrak{G}}(n) = \sum_{t=1}^{n} \lambda(\omega_t, \gamma_t)$$

and the total loss of expert $\mathcal{E}_i$ be

$$\mathrm{Loss}_{\mathcal{E}_i}^{\mathfrak{G}}(n) = \sum_{t=1}^{n} \lambda(\omega_t, \gamma_t^{(i)}) \ ,$$

where $i = 1, 2, \ldots, N$. The upper index $\mathfrak{G}$ can be omitted when it is clear from the context which game we are referring to.

One may think of the predictions $\gamma_t^{(1)}, \gamma_t^{(2)}, \ldots, \gamma_t^{(N)}$ as output by experts $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_N$. Note that the term 'expert' is only a convenient metaphor. In fact we have a full information game between two parties. Our adversary generates predictions $\gamma_t^{(1)}, \gamma_t^{(2)}, \ldots, \gamma_t^{(N)} \in \Gamma$ and outcomes ω_t while we generate predictions γ_t. When we say below that a certain inequality for the total loss of the merging strategy is guaranteed, we mean that it holds no matter what experts' predictions and outcomes are generated by the adversary.

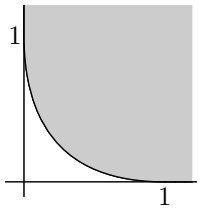

Fig. 1. The square-loss game

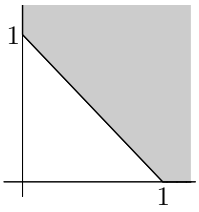

Fig. 2. The absolute-loss game

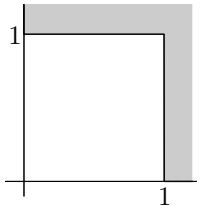

Fig. 3. The simple prediction game

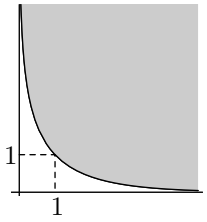

Fig. 4. The logarithmic game

2.3 Geometric Interpretation of a Game

Take a game $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ such that $\Omega = \{\omega^{(0)}, \omega^{(1)}, \ldots, \omega^{(M-1)}\}$ and $|\Omega| = M$. The following important definition goes back to [4,5].

Definition 1 *An M-tuple $(s_0, s_1, \ldots, s_{M-1}) \in [0, +\infty]^M$ is a* superprediction *if there is $\gamma \in \Gamma$ such that the inequalities $\lambda(\omega^{(i)}, \gamma) \leq s_i$ hold for every $i = 0, 1, 2, \ldots, M-1$.*

The set of superpredictions S is an important object characterising the game. Figures 1–4 show the sets of superpredictions for the sample binary games defined in Subsection 2.1.

2.4 Mixability

In this subsection we formulate the result concerning prediction with expert advice for the so called mixable games. It will not be used in our proofs, but it is important for the motivation.

Take a game $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ such that $\Omega = \{\omega^{(0)}, \omega^{(1)}, \ldots, \omega^{(M-1)}\}$ and $|\Omega| = M$. Let $S \subseteq [0, +\infty]^M$ be its set of superpredictions. Take a $\beta \in (0, 1)$ and consider the homeomorphism $\mathfrak{B}_\beta : [0, +\infty]^M \to [0, 1]^M$ specified by the formula $\mathfrak{B}_\beta((x_0, x_1, \ldots, x_{M-1})) = (\beta^{x_0}, \beta^{x_1}, \ldots, \beta^{x_{M-1}})$. We can now give the following definition (after [4,5]).

Definition 2 *We say that $\mathfrak{G}$ is* β-mixable*, where $\beta \in (0, 1)$, if the set $\mathfrak{B}_\beta(S)$ is convex. If $\mathfrak{G}$ is β-mixable for some $\beta \in (0, 1)$, we say that it is* mixable*.*

For mixable games and only for them we can predict as well as the best expert up to an additive constant; the result can be achieved by a merging strategy following the Aggregating Algorithm (AA) [1] .

Proposition 3 ([4,5]) *For every game* $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$

(i) if $\mathfrak{G}$ *is* β*-mixable for some* $\beta \in (0,1)$*, then for every* $N = 1, 2, \ldots$ *and for every merging strategy* $\mathfrak{M}$ *for* N *experts that follows the Aggregating Algorithm, the bound*

$$\mathrm{Loss}_{\mathfrak{M}}(n) \le \mathrm{Loss}_{\mathcal{E}^{(i)}}(n) + \frac{\ln N}{\ln(1/\beta)}$$

is guaranteed for all $n = 1, 2, \ldots$ *and all* $i = 1, 2, \ldots, N$*;*

(ii) if there is a merging strategy $\mathfrak{M}$ *for two experts and a positive constant* a *such that the inequality*

$$\mathrm{Loss}_{\mathfrak{M}}(n) \le \mathrm{Loss}_{\mathcal{E}^{(i)}}(n) + a$$

is guaranteed for all $n = 1, 2, \ldots$ *and* $i = 1, 2$*, then* $\mathfrak{G}$ *is mixable.*

If fact, the results concerning the AA hold for a wider class of games with infinite sets of outcomes Ω. The AA can also be shown to be optimal: the constants it achieves in the upper bounds are optimal.

It can be easily shown directly that the square-loss and the logarithmic games are mixable while the absolute-loss and the simple prediction games are not. This is also implied by more general Lemmas 16 and 17 from Appendix A.

2.5 Convexity

Take a game $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ such that $\Omega = \{\omega^{(0)}, \omega^{(1)}, \ldots, \omega^{(M-1)}\}$ and $|\Omega| = M$. Let $S \subseteq [0, +\infty]^M$ be its set of superpredictions. Let us give the following geometrical definition.

Definition 4 *We say that* $\mathfrak{G}$ *is* convex *if the finite part* $S \cap \mathbb{R}^M$ *of its set of superpredictions* S *is convex.*

Remark 5 *Suppose that* Γ *is a convex set. Then convexity of all the functions* $\lambda(\omega^{(i)}, \gamma)$*,* $i = 0, 1, \ldots, M-1$*, in the second argument implies convexity of the game. However the opposite is not true. Indeed, consider a binary game*

[1] The Aggregating Algorithm as well as the Weak Aggregating Algorithm introduced in this paper leave some flexibility in the choice of the actual predictions; that is the reason why we do not call them merging strategies in the strict sense.

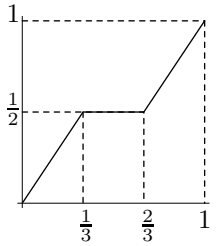

Fig. 5. The loss $\lambda(\omega^{(0)}, \gamma)$ for the example from Remark 5

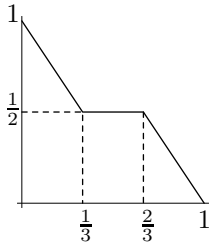

Fig. 6. The loss $\lambda(\omega^{(1)}, \gamma)$ for the example from Remark 5

with $\Gamma = [0, 1]$ and the loss function specified by Figures 5 and 6. Its set of superpredictions coincides with that of the absolute-loss game (Figure 2), but its loss function is not convex in the second argument.

All mixable games are convex, while the opposite is not true. For example, the absolute-loss game is convex but not mixable. A discussion of convexity and mixability and more examples can be found in Appendix A. The simple prediction game provides an example of a non-convex game.

3 Weak Mixability

For non-mixable games it is not possible to predict as well as the best expert up to an additive constant. Let us relax this requirement and ask whether it is possible to predict as well as the best expert up to a larger term.

In the worst case, loss grows linearly in the length of the sequence. Therefore all terms of slower growth can be considered small as compared to loss. This motivates the following definition.

Definition 6 *A game $\mathfrak{G}$ is* weakly mixable *if there is a merging strategy $\mathfrak{M}$ for two experts and a function $f : \mathbb{N} \to \mathbb{R}$ (here $\mathbb{N} = \{1, 2, \ldots\}$ is the set of positive integers) such that $f(n) = o(n)$ as $n \to +\infty$ and the bound*

$$\mathrm{Loss}^{\mathfrak{G}}_{\mathfrak{M}}(n) \le \mathrm{Loss}^{\mathfrak{G}}_{\mathcal{E}^{(i)}}(n) + f(n) \tag{4}$$

is guaranteed for every $n = 1, 2, \ldots$ and $i = 1, 2$.

We can give an equivalent definition requiring that for every $N = 1, 2, 3, \ldots$ there is a merging strategy $\mathfrak{M}$ for N experts such that the inequality

$$\mathrm{Loss}^{\mathfrak{G}}_{\mathfrak{M}}(n) \le \mathrm{Loss}^{\mathfrak{G}}_{\mathcal{E}^{(i)}}(n) + f(n) \ln N \tag{5}$$

is guaranteed for all $n = 1, 2, \ldots$ and for $i = 1, 2, \ldots, N$.

Indeed, a strategy merging two experts can be turned into a strategy merging

N experts by means of the following trick. Let us split the pool of experts into pairs and merge the two experts' predictions inside each pair. Then we can iterate the procedure until we merge all experts' predictions into one. (Note that functions $f(n)$ in the two definitions are different because iterative merging incurs overheads.)

In fact, we shall obtain stronger bounds below. The extra term in the upper bound for the Weak Aggregating Algorithm grows in N as $O(\sqrt{\ln N})$ (see Corollary 14).

The following theorem is the main result of the paper.

Theorem 7 *A game $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ is weakly mixable if and only if it is convex, i.e., the finite part $S \cap \mathbb{R}^M$ of the set of superpredictions S is convex.*

Examples of weakly mixable games are the logarithmic and the square-loss game, which are also mixable, and the absolute-loss game, which is not mixable. The simple prediction game is not weakly mixable.

The rest of the paper contains the proof of the theorem. The 'only if' part follows from Theorem 9 that is formulated in Section 4 and proved in Appendix B.

The 'if' splits into two parts, for bounded and for unbounded games. The 'if' part for bounded games follows from [6]. In Section 5 we shall give an alternative derivation, which achieves a slightly better value of the constant C in the additive term $C\sqrt{n}$. The unbounded case is described in Section 6.

Remark 8 *Let us allow (within this remark) λ to assume negative values; they can be interpreted as 'gain' or 'reward'. If λ assumes the value $-\infty$, the expression for the total loss may include the sum $(-\infty) + (+\infty)$, which is undefined. In order to avoid this ambiguity, it is natural to prohibit λ to take the value $-\infty$. Since λ is assumed to be continuous and Γ compact, this implies that λ is bounded from below, i.e., there is $a > -\infty$ such that $\lambda(\omega, \gamma) \geq a$ for all values of ω and γ.*

Consider another game with the loss function $\lambda'(\omega, \gamma) = \lambda(\omega, \gamma) - a$, which is nonnegative. A merging strategy working with nonnegative loss functions can be easily adapted to work with the original game: let the learner just imagine that it is playing the game with λ'. The losses w.r.t. the two games on a string $\omega_1\omega_2\ldots\omega_n$ will differ by the term an and the upper bounds of the type (4) will be preserved. On the other hand, the sets of superpredictions for the two games will differ by a shift, which preserves convexity. Therefore Theorem 7 remains true for games with loss functions bounded from below.

4 'Only If' Part

We shall derive a statement that is, in fact, slightly stronger than required by Theorem 7.

Theorem 9 *If a game $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$, $|\Omega| = M < +\infty$, has the set of super-predictions S such that its finite part $S \cap \mathbb{R}^M$ is not convex, then there are sequences of experts' predictions $\gamma_t^{(1)} \in \Gamma$ and $\gamma_t^{(2)} \in \Gamma$, $t = 1, 2, \ldots$, and a constant $\theta > 0$ such that for any merging strategy $\mathfrak{S}$ for two experts there is a sequence $\omega_t \in \Omega$, $t = 1, 2, \ldots$, such that*

$$\max_{i=1,2} \left(\mathrm{Loss}_{\mathfrak{S}}^{\mathfrak{G}}(n) - \mathrm{Loss}_{\mathcal{E}_i}^{\mathfrak{G}}(n) \right) \geq \theta n \tag{6}$$

for all positive integers n.

For the proof see Appendix B.

5 'If' Part for Bounded Games

5.1 Weak Aggregating Algorithm

In this subsection we introduce the Weak Aggregating Algorithm (WAA). Let $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ be a game such that $|\Omega| = M < +\infty$ and let N be the number of experts. Let $\Omega = \{\omega^{(0)}, \omega^{(1)}, \ldots, \omega^{(M-1)}\}$. Although the WAA can be applied to any game, the performance results we shall obtain hold only for bounded games, so one may assume that $\mathfrak{G}$ is bounded.

We describe the WAA using pseudo-code. The WAA accepts N initial normalised weights $q_1, q_2, \ldots, q_N \in [0, 1]$ such that $\sum_{i=1}^N q_i = 1$ and a positive number c as parameters. The role of c is similar to that of the learning rate in the theory of prediction with expert advice. Let $\beta_t = e^{-c/\sqrt{t}}$, $t = 1, 2, \ldots$

(1) $l_1^{(i)} := 0$, $i = 1, 2, \ldots, N$
(2) FOR $t = 1, 2, \ldots$
(3) $\quad w_t^{(i)} := q_i \beta_t^{l_t^{(i)}}$, $i = 1, 2, \ldots, N$
(4) $\quad p_t^{(i)} := \frac{w_t^{(i)}}{\sum_{j=1}^N w_t^{(j)}}$, $i = 1, 2, \ldots, N$
(5) $\quad$ read experts' predictions $\gamma_t^{(1)}, \gamma_t^{(2)}, \ldots, \gamma_t^{(N)}$
(6) $\quad g_k := \sum_{j=1}^N \lambda\left(\omega^{(k)}, \gamma_t^{(j)}\right) p_t^{(j)}$, $k = 0, 1, \ldots, M-1$
(7) $\quad$ output $\gamma_t \in \Gamma$ such that $\lambda(\omega^{(k)}, \gamma_t) \leq g_k$ for all $k = 0, 1, \ldots, M-1$

(8) observe ω_t
(9) $l_{t+1}^{(i)} := l_t^{(i)} + \lambda\left(\omega_t, \gamma_t^{(i)}\right)$, $i = 1, 2, \ldots, N$
(10) END FOR

The variable $l_t^{(i)}$ stores the loss of the i-th expert $\mathcal{E}^{(i)}$, i.e., after trial t we have $l_{t+1}^{(i)} = \mathrm{Loss}_{\mathcal{E}^{(i)}}^{\mathfrak{G}}(t)$. The values $w_t^{(i)}$ are weights assigned to the experts during the work of the algorithm; they depend on the loss suffered by the experts and the initial weights q_i. The values $p_t^{(i)}$ are obtained by normalising $w_t^{(i)}$. Note that from the computational point of view it is sufficient to have only one set of variables $p^{(i)}$, $i = 1, 2, \ldots, N$, one set of variables $w^{(i)}$, $i = 1, 2, \ldots, N$, and one set of variables $l^{(i)}$, $i = 1, 2, \ldots, N$ to save memory. The subscript t has been added in order to simplify referring to these variables in the proofs below.

This algorithm is applicable if the set of superpredictions S has a convex finite part $S \cap \mathbb{R}^M$. If this is the case, then the point $(g_0, g_1, \ldots, g_{M-1})$ belongs to S and thus γ_t can be found on step (7). The choice of γ_t is not necessarily unique.

Remark 10 *Suppose that Γ is a convex set and the functions $\lambda(\omega, \gamma)$ are convex in the second argument for all $\omega \in \Omega$ (cf. Remark 5). Then on step (7) we can take $\gamma_t = \sum_{j=1}^{N} p_t^{(j)} \gamma_t^{(j)}$ (note that it is not necessarily the only possible solution).*

For bounded games the following lemma holds.

Lemma 11 *For every $L > 0$, every game $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ such that $|\Omega| < +\infty$ and $\lambda(\omega, \gamma) \le L$ for all $\omega \in \Omega$ and $\gamma \in \Gamma$ and every $N = 1, 2, \ldots$, for every merging strategy $\mathfrak{M}$ for N experts that follows the WAA with initial weights $q_1, q_2, \ldots, q_N \in [0, 1]$ such that $\sum_{i=1}^{N} q_i = 1$ and $c > 0$ the bound*

$$\mathrm{Loss}_{\mathfrak{M}}(n) \le \mathrm{Loss}_{\mathcal{E}^{(i)}}(n) + \left(cL^2 + \frac{1}{c} \ln \frac{1}{q_i}\right) \sqrt{n} \tag{7}$$

is guaranteed for every $n = 1, 2, \ldots$ and every $i = 1, 2, \ldots, N$.

The proof of Lemma 11 is given in Appendix C.

Remark 12 *It is easy to see that the result of Lemma 11 will still hold for a countable pool of experts $\mathcal{E}_1, \mathcal{E}_2, \ldots$ We take weights $\sum_{i=1}^{+\infty} q_i = 1$; the sums in lines (4) and (6) from the definition of the WAA become infinite but they clearly converge. The point $(g_0, g_1, \ldots, g_{M-1})$ clearly belong to S because S is closed (in fact, convexity is sufficient here; a convex combination of countably many points still belongs to their convex closure; see e.g., Theorem 2.4.1 in [8]).*

Remark 13 *The WAA belongs to the class of merging strategies that on each step produce a distribution* $p_t^{(1)}, p_t^{(2)}, \ldots, p_t^{(N)}$ *and suffer loss bounded by the weighted sum of experts' losses* $\sum_{i=1}^N \lambda(\omega_t, \gamma_t^{(i)}) p_t^{(i)}$. *This means that WAA can be applied to* every *bounded game in the following randomised fashion. Let us choose one expert from the pool randomly according to this distribution and output the prediction of that expert. Our expected loss on each step will be bounded by the same weighted sum. Therefore (7) will hold with the left-hand side replaced by* $\mathbf{E}\,\mathrm{Loss}_{\mathfrak{M}}(n)$, *where the expectation* $\mathbf{E}$ *is taken w.r.t. the internal randomisation. Note that the convexity requirement becomes unnecessary; introducing the randomisation has essentially the same effect as taking the convex hull of the set of superpredictions.*

Let us take equal initial weights $q_1 = q_2 = \ldots = q_N = 1/N$ in the WAA. The additive term then reduces to $(cL^2 + (\ln N)/c)\sqrt{n}$. When $c = \sqrt{\ln N}/L$, this expression reaches its minimum. We get the following corollary.

Corollary 14 *Under the conditions of Lemma 11, there is a merging strategy* $\mathfrak{M}$ *such that the bound*

$$\mathrm{Loss}_{\mathfrak{M}}(n) \le \mathrm{Loss}_{\mathcal{E}^{(i)}}(n) + 2L\sqrt{n \ln N}$$

is guaranteed.

6 'If' Part for Unbounded Games

6.1 Counterexample

The WAA can be applied even in the case of an unbounded game; indeed, the only requirement is that the finite part of the set of superpredictions S is convex. However we cannot guarantee that a reasonable upper bound on the loss of a strategy that uses it will exist. The same applies to any strategy that uses a linear combination in the same fashion as WAA.

Indeed, consider a game with an unbounded loss function λ. Let ω_0 be such that the function $\lambda(\omega_0, \gamma)$ attains arbitrary large values.

Suppose that there are two experts $\mathcal{E}_1$ and $\mathcal{E}_2$ and on some trial they are ascribed weights $p^{(1)}$ and $p^{(2)}$ such that $p^{(2)} > 0$. Suppose that $\mathcal{E}_1$ outputs $\gamma^{(1)}$ such that $\lambda(\omega_0, \gamma^{(1)}) < +\infty$ (see Figure 7 for a two-dimensional illustration). The upper estimate on the loss of the merging strategy in the case when the outcome ω_0 occurs is

$$g_0 = p^{(1)}\lambda(\omega_0, \gamma^{(1)}) + p^{(2)}\lambda(\omega_0, \gamma^{(2)}) \ ,$$

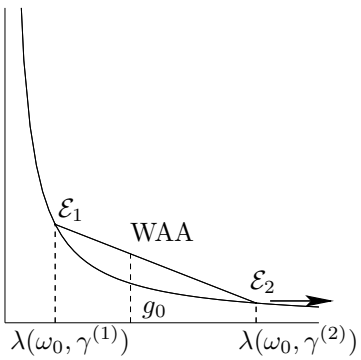

Fig. 7. A counterexample for unbounded games in dimension 2.

where $\gamma^{(2)}$ is the prediction output by $\mathcal{E}_2$. Let us vary $\gamma^{(2)}$. The weights depend on the previous behaviour of the experts and they cannot be changed. If $\lambda(\omega_0, \gamma^{(2)})$ tends to infinity, then g_0 tends to infinity and therefore the difference $g_0 - \lambda(\omega_0, \gamma^{(1)})$ tends to infinity. Thus the learner cannot compete with the first expert.

This example shows that the WAA cannot be straightforwardly generalised to unbounded games. It needs to be altered.

6.2 Approximating Unbounded Games with Bounded

The following lemma allows us to 'cut off' the infinity at a small cost.

Lemma 15 *Let $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ be a game such that $|\Omega| < +\infty$. Then for every $\varepsilon > 0$ there is $L_\varepsilon > 0$ with the following property. For every $\gamma \in \Gamma$ there is $\gamma^* \in \Gamma$ such that $\lambda(\omega, \gamma^*) \le L_\varepsilon$ and $\lambda(\omega, \gamma^*) \le \lambda(\omega, \gamma) + \varepsilon$ for all $\omega \in \Omega$.*

The proof of Lemma 15 is given in Appendix D.

In the case of two outcomes $|\Omega| = 2$ obtaining L_ε is particularly straightforward. See Figure 8, where

$$C = \inf_{\gamma \in \Gamma} \lambda(\omega^{(0)}, \gamma) \text{ and } D = \inf_{\gamma \in \Gamma} \lambda(\omega^{(1)}, \gamma) \ ;$$

we can take $L_\varepsilon = \max(L_0, L_1)$. If γ is such that the point $(\lambda(\omega^{(0)}, \gamma), \lambda(\omega^{(1)}, \gamma))$ falls into the area to the right of the straight line $x = L_0$, we can take γ^* such that $(\lambda(\omega^{(0)}, \gamma^*), \lambda(\omega^{(1)}, \gamma^*)) = (L_0, D + \varepsilon)$.

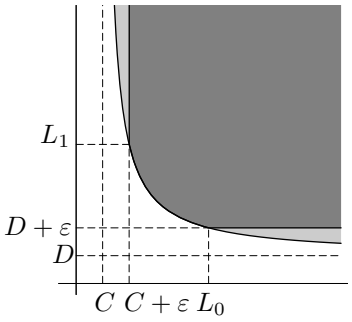

Fig. 8. Obtaining L_ε in the case of two outcomes.

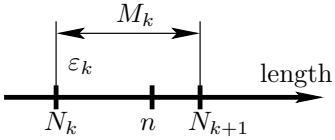

Fig. 9. The sequences of N_k, M_k, and ε_k.

6.3 Merging Experts in the Unbounded Case

Consider an unbounded game $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ and N experts $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_N$. Fix some $\varepsilon > 0$. Let L_ε be as above. After obtaining experts' predictions $\gamma_t^{(1)}, \gamma_t^{(2)}, \ldots, \gamma_t^{(N)}$ we can find $\gamma_t^{(1)*}, \gamma_t^{(2)*}, \ldots, \gamma_t^{(N)*}$ as in Lemma 15 and then apply the results from the bounded case to them. By proceeding in this fashion, a strategy $\mathfrak{M}$ suffers loss such that

$$\mathrm{Loss}_{\mathfrak{M}}^{\mathfrak{G}}(n) \le \mathrm{Loss}_{\mathcal{E}^{(i)}}^{\mathfrak{G}}(n) + C_\varepsilon \sqrt{n} + \varepsilon n \tag{8}$$

for all $i = 1, 2, \ldots, N$ and $\omega_1, \omega_2, \ldots, \omega_n \in \Omega$, $n = 1, 2, \ldots$, where $C_\varepsilon = 2L_\varepsilon^2 \sqrt{\ln N}$ (we are applying WAA with equal weights).

This inequality does not allow us to prove Theorem 7. In order to achieve an extra term of the order $o(n)$ we shall vary ε.

Take a strictly increasing sequence of integers N_k, $k = 1, 2, \ldots$, and a sequence $\varepsilon_k > 0$, $k = 0, 1, 2, \ldots$ Consider the merging strategy $\mathfrak{M}$ defined as follows. The strategy first takes ε_0 and merges the experts' predictions using the WAA and ε_0 in the fashion described above. This continues while n, the length of the sequence of outcomes, is less than or equal to N_1. Then the strategy switches to ε_1 and applies the WAA and ε_1 until n exceeds N_2 etc (see Figure 9). Note that each time n passes through a limit N_i, the current invocation of the WAA terminates and a completely new invocation of the WAA starts working. It does not have to inherit anything from previous invocations.

In Appendix E we show how to choose the sequences ε_k and N_k in such a way as to achieve the desired extra term of the order $o(n)$.

7 Computability Issues

Since the results of this paper are proved constructively, they can be restated in a constructive fashion.

Let us require in Definition 6 that $\mathfrak{M}$ is computable. The experts do not have to be computable in any sense because in our analysis the merging strategy has no access to their internal 'machinery'. The merging strategy simply receives experts' predictions as inputs. Note that we can choose computable sequences $\gamma_t^{(1)}$ and $\gamma_t^{(2)}$ in Theorem 9 though. The sequence ω_n can be generated effectively if $\mathfrak{M}$ is computable.

In order for the merging strategies constructed in the proof of Theorem 7 to be computable, we need to impose computability restrictions on games. We require the loss function to be computable so that the operations we need to do become possible.

We need to be able do the following. First we need to compute the values of λ. Secondly in order to perform step (7) of the WAA we need to be able to solve systems of inequalities of the type

$$\begin{aligned} \lambda(\omega^{(0)}, \gamma) &\le t_0 \ , \\ \lambda(\omega^{(1)}, \gamma) &\le t_1 \ , \\ &\dots \\ \lambda(\omega^{(M-1)}, \gamma) &\le t_{M-1} \end{aligned}$$

w.r.t. γ, where $t_i = \sum_{j=1}^{m} p_j \lambda(\omega^{(i)}, \gamma_j)$ for some set of γ_j and weights p_j ($i = 0, 1, \dots, M-1$ and $j = 1, 2, \dots, N$). Note that we only encounter systems where the solution is known to exist. Thirdly for unbounded games we need to compute the values L_ε from Lemma 15. If we have the value of L_ε, we can find γ^* for every γ_0 by solving the system of the aforementioned type with $t_i = \min(\lambda(\omega^{(i)}, \gamma_0) + \varepsilon, L_\varepsilon)$, $i = 0, 1, \dots, M-1$.

These requirements are quite natural and every reasonable loss function (e.g., specified by a reasonable analytical expression) should satisfy them.

Remark 10 simplifies our task if Γ is convex and λ convex in the second argument. We can then find γ_t on step (7) of the WAA by taking a convex combination on Γ.

Suppose that we have an oracle that can answer the questions of the types we have listed. Then both the WAA and the algorithm for unbounded functions we have constructed output the prediction on each step of the on-line protocol in $O(MN)$ time modulo calls to the oracle.

Appendix A. Convexity vs Mixability

In this appendix we show that convexity is a weaker requirement than mixability. All mixable games are convex, while the converse is not true. We shall give a geometrical proof and construct examples.

Lemma 16 *If a game $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ such that $\Omega = \{\omega^{(0)}, \omega^{(1)}, \ldots, \omega^{(M-1)}\}$ is mixable, then it is convex.*

PROOF. We shall rely on a characterisation of convexity by means of support hyperplanes (see, e.g., Theorems 8 and 9 in [9]).

Take a point $x_0 = \left(x_0^{(0)}, x_0^{(1)}, \ldots, x_0^{(M-1)}\right) \in \mathbb{R}^M$ on the boundary $\partial(S \cap \mathbb{R}^M)$. Let $\beta \in (0,1)$ be such that $\mathfrak{G}$ is β-mixable and hence $\mathfrak{B}_\beta(S)$ is convex. Through the point $\mathfrak{B}_\beta(x_0)$ there passes a support hyperplane to $\mathfrak{B}_\beta(S)$.

Because the set $\mathfrak{B}_\beta(S)$ contains the whole parallelepiped with the diagonal from the origin to $\mathfrak{B}_\beta(x_0)$, the equation of the hyperplane can be written as $\sum_{i=0}^{M-1} a_i u^{(i)} = 1$, where $a_i \geq 0$ for all $i = 0, 1, \ldots, M-1$ (here $u^{(i)}$ are the coordinates in $\mathbb{R}^M$).

Therefore the set S lies 'above' the surface passing through x_0 and specified by the equation $\sum_{i=0}^{M-1} a_i \beta^{x^{(i)}} = 1$ (here $x^{(i)}$ are the coordinates in $\mathbb{R}^M$), where $a_i \geq 0$ for all $i = 0, 1, \ldots, M-1$. Since $a\beta^x = \beta^{x + \log_\beta a}$, this surface is a shift of either the surface $\sum_{i=0}^{M-1} \beta^{x^{(i)}} = 1$ or a cylinder over a similar surface of lower dimension. Since the function β^x is concave, the sum $\sum_{i=0}^{M-1} \beta^{x^{(i)}}$ is concave and the set $\left\{\left(x^{(0)}, x^{(1)}, \ldots, x^{(M-1)}\right) \in \mathbb{R}^{M-1} \mid \beta^{x^{(i)}} \leq 1\right\}$ is convex. A support hyperplane passes through each point on the surface; thus we can draw a support hyperplane to $S \cap \mathbb{R}^M$ through x_0. $\square$

We shall now construct binary examples differentiating convex games from mixable. We need the following lemma from [10] (it is in fact a restatement of results from [2]).

Lemma 17 *Let $\mathfrak{G}$ be a binary game with the set of superpredictions S. Suppose that there are twice differentiable functions $x, y : I \to \mathbb{R}$, where $I \subseteq \mathbb{R}$ is an open (perhaps infinite) interval, such that $x' > 0$ and $y' < 0$ on I and S is the closure of the set $\{(u, v) \in \mathbb{R}^2 \mid$ there is $t \in I : x(t) \leq u$ and $y(t) \leq v\}$ w.r.t. the extended topology of $[-\infty, +\infty]^2$. Then, for every $\beta \in (0,1)$, the game $\mathfrak{G}$ is β-mixable if and only if*

$$\ln \frac{1}{\beta} \leq \frac{y''(t)x'(t) - x''(t)y'(t)}{x'(t)y'(t)(y'(t) - x'(t))}$$

holds for every $t \in I$. The game $\mathfrak{G}$ is mixable if and only if the fraction $(y''x' - x''y')/x'y'(y' - x')$ is separated from the zero, i.e., there is $\varepsilon > 0$ such that

$$\frac{y''x' - x''y'}{x'y'(y' - x')} \geq \varepsilon \tag{9}$$

holds on I.

PROOF. Convexity of $\mathfrak{B}_\beta(S)$ is equivalent to concavity of the function with the graph $\{\mathfrak{B}_\beta(x(t), y(t)) \mid t \in I\}$. Because the functions $x(t)$ and $y(t)$ are smooth, this curve is concave if and only if the inequality

$$\frac{d^2\beta^{y(t)}}{d\left(\beta^{x(t)}\right)^2} \leq 0$$

holds on I. Differentiation yields

$$\frac{d\beta^{y(t)}}{d\beta^{x(t)}} = \beta^{y(t)-x(t)}\frac{y'(t)}{x'(t)}$$

and

$$\frac{d^2\beta^{y(t)}}{d\left(\beta^{x(t)}\right)^2} = \frac{\beta^{y(t)-2x(t)}}{\ln\beta \cdot (x'(t))^2}\left((y'(t) - x'(t))y'(t)\ln\beta + \frac{y''(t)x'(t) - y'(t)x''(t)}{x'(t)}\right) .$$

The lemma follows. □

Using this lemma, one can check that the square-loss and the logarithmic games are mixable, while the absolute-loss game is not.

If in the lemma $x(t) = t$, one can rewrite (9) as

$$\frac{y''}{y'(y' - 1)} \geq \varepsilon > 0 .$$

The convexity requirement reduces to $y'' \geq 0$. These formulae allow us to construct various examples of convex games that are not mixable.

If the second derivative of $y(x)$ vanishes inside the interval (but $y(x)$ does not become constant), then $y(x)$ specifies the set of superpredictions of a non-mixable game.

The following group of examples shows that mixability can be violated 'at the infinity'. Let $I = (0, +\infty)$ and $y(x) = 1/x^m$, $m > 0$. We have

$$\frac{y''}{y'(y' - 1)} = \frac{(m + 1)x^m}{m + x^{m+1}}$$

and the fraction tends to 0 as $x \to 0$ or $x \to +\infty$. Clearly all the games with the sets of superpredictions specified by such $y(x)$ are convex and unbounded but not mixable. However if we cut off the ends of the interval and take $I = (a, b)$, where $0 < a < b < +\infty$, we get mixable games.

Appendix B. Proof of the 'Only If' Part

PROOF of Theorem 9.

We shall use the following simple vector notation. If $X = (x_1, \ldots, x_n)$, $Y = (y_1, \ldots, y_n)$ and $\alpha \in \mathbb{R}$, then $X + Y$ and αX are defined in the natural way. By $\langle X, Y \rangle$ we denote the scalar product $\sum_{i=1}^{n} x_i y_i$. Vector inequalities, e.g., $X \geq Y$, hold if they hold component-wise. Note that the definition of the set of superpredictions S implies that if $X \in S$ and $Y \geq X$ than $Y \in S$.

For brevity we shall denote finite sequences by bold letters, e.g., $\boldsymbol{x} = \omega_1 ... \omega_n \in \Omega^n$. Let $|\boldsymbol{x}|$ be the length of $\boldsymbol{x}$, i.e., the total number of symbols in $\boldsymbol{x}$. We shall denote the number of elements equal to $\omega^{(0)}$ in a sequence $\boldsymbol{x}$ by $\sharp_0 \boldsymbol{x}$, the number of elements equal to $\omega^{(1)}$ by $\sharp_1 \boldsymbol{x}$ etc. It is easy to see that $\sum_{i=0}^{M-1} \sharp_i \boldsymbol{x} = |\boldsymbol{x}|$ for every $\boldsymbol{x} \in \Omega^*$. The vector $(\sharp_0 \boldsymbol{x}, \sharp_1 \boldsymbol{x}, \ldots, \sharp_{M-1} \boldsymbol{x})$ will be denoted by $\sharp \boldsymbol{x}$.

There are points $B_1 = \left(b_1^{(0)}, b_1^{(1)}, \ldots, b_1^{(M-1)}\right)$ and $B_2 = \left(b_2^{(0)}, b_2^{(1)}, \ldots, b_2^{(M-1)}\right)$ such that $B_1, B_2 \in S \cap \mathbb{R}^M$ but the segment $[B_1, B_2]$ connecting them is not a subset of S. Let $\alpha \in (0, 1)$ be such that $C = \alpha B_1 + (1 - \alpha) B_2$ does not belong to S (see Figure 10). Since λ is continuous and Γ is compact, the set S is closed and thus there is a small vicinity of C that is a subset of $\mathbb{R}^M \setminus S$.

Without restricting the generality one may assume that all coordinates of B_1 and B_2 are strictly positive. Indeed, the points $B_1' = B_1 + t \cdot (1, 1, \ldots, 1)$ and $B_2' = B_2 + t \cdot (1, 1, \ldots, 1)$ belong to S for all positive t. If $t > 0$ is sufficiently small, then $C' = \alpha B_1' + (1 - \alpha) B_2'$ still belongs to the vicinity mentioned above and thus C' does not belong to S.

Let us draw a straight line l through the origin and point C. Let $A = \left(a^{(0)}, a^{(1)}, \ldots, a^{(M-1)}\right)$ be the intersection of l with the boundary ∂S. Such a point really exists. Indeed, $l = \{X \in \mathbb{R}^M \mid \exists t \geq 0 \ : \ X = tC\}$. For sufficiently large t all coordinates of tC are greater than the corresponding coordinates of B_1 and thus $tC \in S$. Now let $t_0 = \inf\{t \geq 0 \mid tC \in S\}$ and $A = t_0 C$. Since $C \notin S$, we get $t_0 > 1$ and thus $A = (1 + \delta) C$, where $\delta > 0$.

We now proceed to constructing the sequences $\gamma_t^{(1)}$ and $\gamma_t^{(2)}$. There are predictions $\gamma_1, \gamma_2 \in \Gamma$ such that $\lambda(\omega^{(i)}, \gamma_1) \leq b_1^{(i)}$ and $\lambda(\omega^{(i)}, \gamma_2) \leq b_2^{(i)}$ for all

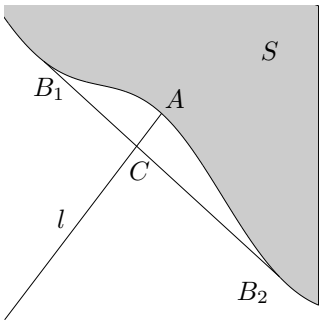

Fig. 10. The drawing for the proof of Theorem 9.

$i = 0, 1, 2, \ldots, M-1$. Let $\gamma_t^{(1)} = \gamma_1$ and $\gamma_t^{(2)} = \gamma_2$ for all $t = 1, 2, \ldots$ If $\boldsymbol{x} = \omega_1\omega_2\ldots,\omega_t$, then

$$\mathrm{Loss}_{\mathcal{E}_1}(t) \le \sum_{i=0}^{M-1} \sharp_i \boldsymbol{x} b_1^{(i)} = \langle B_1, \sharp\boldsymbol{x}\rangle \ , \tag{10}$$

$$\mathrm{Loss}_{\mathcal{E}_2}(t) \le \sum_{i=0}^{M-1} \sharp_i \boldsymbol{x} b_2^{(i)} = \langle B_2, \sharp\boldsymbol{x}\rangle \tag{11}$$

for all $t = 1, 2, \ldots$

Now let us consider a merging strategy $\mathfrak{S}$ and construct a sequence $\boldsymbol{x}_n = \omega_1\omega_2\ldots\omega_n$ satisfying the requirements of the theorem. The sequence is constructed by induction. Suppose that $\boldsymbol{x}_n$ has been constructed. Let γ be the prediction output by $\mathfrak{S}$ on the $(n+1)$-th trial, provided the previous outcomes were elements constituting the strings $\boldsymbol{x}_n$ in the correct order. There is some $\omega^{(i_0)} \in \Omega$ such that $\lambda(\omega^{(i_0)}, \gamma) \ge a^{(i_0)}$. Indeed, if this is not true and the inequalities $\lambda(\omega^{(i)}, \gamma) < a^{(i)}$ hold for all $i = 1, 2, \ldots, M-1$, then there is a vicinity of A that is a subset of S. This contradicts the definition of A. We let $\boldsymbol{x}_{n+1} = \boldsymbol{x}_n\omega_{i_0}$. The construction implies

$$\mathrm{Loss}_{\mathfrak{S}}(n) \ge \sum_{i=0}^{M-1} \sharp_i \boldsymbol{x}_n a^{(i)} = \langle A, \sharp\boldsymbol{x}_n\rangle \ . \tag{12}$$

Let $\varepsilon = \min_{j=1,2;\ i=0,1,2,\ldots,M-1} b_j^{(i)} > 0$. We get $\langle B_j, \boldsymbol{x}\rangle = \sum_{i=0}^{M-1} b_j^{(i)} \sharp_i \boldsymbol{x} \ge \varepsilon|\boldsymbol{x}|$ for all strings $\boldsymbol{x} \in \Omega^*$ and $j = 1, 2$. Since $A = (1+\delta)(\alpha B_1 + (1-\alpha)B_2)$ we get

$$\begin{aligned}\langle A, \sharp\boldsymbol{x}\rangle &= (1+\delta)(\alpha\langle B_1, \sharp\boldsymbol{x}\rangle + (1-\alpha)\langle B_2, \sharp\boldsymbol{x}\rangle) \\ &\ge \alpha\langle B_1, \sharp\boldsymbol{x}\rangle + (1-\alpha)\langle B_2, \sharp\boldsymbol{x}\rangle + \delta\varepsilon|\boldsymbol{x}|\end{aligned}$$

for all strings $\boldsymbol{x}$. Let $\theta = \delta\varepsilon$; note that ε and δ do not depend on $\mathfrak{S}$. By combining this inequality with (10), (11), and (12) we obtain the inequality

$$\mathrm{Loss}_{\mathfrak{S}}(n) \ge \alpha\,\mathrm{Loss}_{\mathcal{E}_1}(n) + (1-\alpha)\,\mathrm{Loss}_{\mathcal{E}_2}(n) + \theta n$$

for all positive integers n.

It is easy to see that

$$\begin{aligned}\mathrm{Loss}_{\mathfrak{S}}(n) - \mathrm{Loss}_{\mathcal{E}_1}(n) &\geq (1-\alpha)(\mathrm{Loss}_{\mathcal{E}_2}(n) - \mathrm{Loss}_{\mathcal{E}_1}(n)) + \theta n \ , \\ \mathrm{Loss}_{\mathfrak{S}}(n) - \mathrm{Loss}_{\mathcal{E}_2}(n) &\geq \alpha(\mathrm{Loss}_{\mathcal{E}_1}(n) - \mathrm{Loss}_{\mathcal{E}_2}(n)) + \theta n \ .\end{aligned}$$

If $\mathrm{Loss}_{\mathcal{E}_2}(n) \geq \mathrm{Loss}_{\mathcal{E}_1}(n)$ the former difference is greater than or equal to θn, otherwise the latter difference is greater than or equal to θn. By combining these two inequalities we obtain (6). $\Box$

Appendix C. Proof of Lemma 11

In this appendix we prove Lemma 11. We start with the following lemma.

Lemma 18 *Let $\mathfrak{G} = \langle \Omega, \Gamma, \lambda \rangle$ be a game such that $|\Omega| < +\infty$ and let N be the number of experts. Let the finite part of the set of superpredictions $S \cap \mathbb{R}^M$ be convex. If $\mathfrak{M}$ is a merging strategy following the WAA, then for every $t = 1, 2, \ldots$ we get*

$$\beta_t^{\mathrm{Loss}_{\mathfrak{M}}^{\mathfrak{G}}(t)} \geq \beta_t^{\sum_{j=1}^{t} \delta(j)} \sum_{i=1}^{N} q_i \beta_t^{\mathrm{Loss}_{\mathcal{E}^{(i)}}^{\mathfrak{G}}(t)} \ , \tag{13}$$

where

$$\delta(j) = \log_{\beta_j} \frac{\beta_j^{\sum_{i=1}^{N} \lambda(\omega_j, \gamma_j^{(i)}) p_j^{(i)}}}{\sum_{i=1}^{N} \beta_j^{\lambda(\omega_j, \gamma_j^{(i)})} p_j^{(i)}} \tag{14}$$

for $j = 1, 2, \ldots, t$, in the notation introduced above.

PROOF of Lemma 18. The proof is by induction on t. Let us assume that (13) holds and then derive the corresponding inequality for the step $t+1$.

The function x^α, where $0 < \alpha < 1$ and $x \geq 0$, is increasing in x and it is also concave in x. For every set of weights $p_i \in [0,1]$, $i = 1, \ldots, n$ such that $\sum_{i=1}^{n} p_i = 1$ and every array of $x_i \geq 0$, $i = 1, \ldots, n$, we get $(\sum_{i=1}^{n} p_i x_i)^\alpha \geq \sum_{i=1}^{n} p_i x_i^\alpha$.

Therefore (13) implies

$$\beta_{t+1}^{\mathrm{Loss}_{\mathfrak{M}}^{\mathfrak{G}}(t)} = \left(\beta_t^{\mathrm{Loss}_{\mathfrak{M}}^{\mathfrak{G}}(t)}\right)^{\log_{\beta_t}\beta_{t+1}} \tag{15}$$

$$\geq \left(\beta_t^{\sum_{j=1}^{t}\delta(j)} \sum_{i=1}^{N} q_i \beta_t^{\mathrm{Loss}_{\mathcal{E}^{(i)}}^{\mathfrak{G}}(t)}\right)^{\log_{\beta_t}\beta_{t+1}} \tag{16}$$

$$\geq \beta_{t+1}^{\sum_{j=1}^{t}\delta(j)} \sum_{i=1}^{N} q_i \beta_{t+1}^{\mathrm{Loss}_{\mathcal{E}^{(i)}}^{\mathfrak{G}}(t)} \tag{17}$$

Step (7) of the algorithm implies that $\lambda(\omega_{t+1},\gamma_{t+1}) \leq \sum_{i=1}^{N} \lambda\left(\omega_{t+1},\gamma_{t+1}^{(i)}\right) p_{t+1}^{(i)}$. By exponentiating this inequality we get

$$\beta_{t+1}^{\lambda(\omega_{t+1},\gamma_{t+1})} \geq \beta_{t+1}^{\sum_{i=1}^{N}\lambda\left(\omega_{t+1},\gamma_{t+1}^{(i)}\right)p_{t+1}^{(i)}} \tag{18}$$

$$= \frac{\beta_{t+1}^{\sum_{i=1}^{N}\lambda\left(\omega_{t+1},\gamma_{t+1}^{(i)}\right)p_{t+1}^{(i)}}}{\sum_{i=1}^{N}\beta_{t+1}^{\lambda\left(\omega_{t+1},\gamma_{t+1}^{(i)}\right)}p_{t+1}^{(i)}} \sum_{i=1}^{N} \beta_{t+1}^{\lambda\left(\omega_{t+1},\gamma_{t+1}^{(i)}\right)} p_{t+1}^{(i)} \tag{19}$$

$$= \beta_{t+1}^{\delta(t+1)} \sum_{i=1}^{N} \beta_{t+1}^{\lambda\left(\omega_{t+1},\gamma_{t+1}^{(i)}\right)} p_{t+1}^{(i)} \ . \tag{20}$$

Multiplying (17) by (20) and substituting

$$p_{t+1}^{(i)} = \frac{w_{t+1}}{\sum_{j=1}^{N} w_{t+1}^{(j)}} = \frac{q_i \beta_{t+1}^{\mathrm{Loss}_{\mathcal{E}^{(i)}}^{\mathfrak{G}}(t)}}{\sum_{j=1}^{N} q_j \beta_{t+1}^{\mathrm{Loss}_{\mathcal{E}^{(j)}}^{\mathfrak{G}}(t)}}$$

completes the proof on the lemma. □

By taking the logarithm of (13) we get

$$\mathrm{Loss}_{\mathfrak{M}}^{\mathfrak{G}}(t) \leq \sum_{j=1}^{t}\delta(j) + \log_{\beta_t}\sum_{i=1}^{N} q_i \beta_t^{\mathrm{Loss}_{\mathcal{E}^{(i)}}^{\mathfrak{G}}(t)}$$

$$\leq \sum_{j=1}^{t}\delta(j) + \log_{\beta_t} q_i + \mathrm{Loss}_{\mathcal{E}^{(i)}}^{\mathfrak{G}}(t)$$

for every $i = 1, 2, \ldots, N$. We have $\log_{\beta_t} q_i = -\frac{\sqrt{t}}{c}\ln q_i$. It remains to estimate the first term.

Recall that L is an upper bound on λ. By applying the inequality $\ln x \leq x - 1$

we get

$$\begin{aligned}\delta(t) &= \sum_{i=1}^{N} \lambda(\omega_t, \gamma_t^{(i)}) p_t^{(i)} + \frac{\sqrt{t}}{c} \ln \sum_{i=1}^{N} \beta_t^{\lambda(\omega_t, \gamma_t^{(i)})} p_t^{(i)} \\ &\le \sum_{i=1}^{N} \lambda(\omega_t, \gamma_t^{(i)}) p_t^{(i)} + \frac{\sqrt{t}}{c} \left(\sum_{i=1}^{N} \beta_t^{\lambda(\omega_t, \gamma_t^{(i)})} p_t^{(i)} - 1 \right)\end{aligned}$$

By using Taylor's series with Lagrange's remainder term we obtain

$$\beta_t^{\lambda(\omega_t, \gamma_t^{(i)})} = e^{-c\lambda(\omega_t, \gamma_t^{(i)})/\sqrt{t}} = 1 - \frac{c\lambda(\omega_t, \gamma_t^{(i)})}{\sqrt{t}} + \frac{1}{2} \left(\frac{c\lambda(\omega_t, \gamma_t^{(i)})}{\sqrt{t}} \right)^2 e^{\xi} \ ,$$

where $\xi \in [-c\lambda(\omega_t, \gamma_t^{(i)})/\sqrt{t}, 0]$ and thus

$$\beta_t^{\lambda(\omega_t, \gamma_t^{(i)})} \le 1 - \frac{c\lambda(\omega_t, \gamma_t^{(i)})}{\sqrt{t}} + \frac{c^2 L^2}{2t} \ .$$

Therefore $\delta(t) \le cL^2/2\sqrt{t}$ and summation yields

$$\sum_{j=1}^{t} \delta(j) \le \sum_{j=1}^{t} \frac{cL^2}{2\sqrt{j}} \le \frac{cL^2}{2} \int_0^t \frac{dx}{\sqrt{x}} = cL^2\sqrt{t} \ .$$

This completes the proof.

Remark 19 *Let us discuss the intuitive meaning of the term $\delta(t)$. We have*

$$\delta(t) = \sum_{i=1}^{N} \lambda(\omega_t, \gamma_t^{(i)}) p_t^{(i)} - \log_{\beta_t} \left(\sum_{i=1}^{N} \beta_t^{\lambda(\omega_t, \gamma_t^{(i)})} p_t^{(i)} \right) \ .$$

This is the difference of two terms corresponding to two different ways of mixing experts' predictions. The first is the convex mixture we use in the WAA. The second is the mixture used in the Aggregating Algorithm (AA) (see [4,5]). In the AA the transformation $\mathfrak{B}_\beta$ (see Subsection 2.4) is applied, a mixture is calculated in the image space, and then the inverse image $\mathfrak{B}_\beta^{-1}$ is taken. This is only possible if the game is β-mixable, while for the WAA convexity is sufficient. What we have shown is that the loss suffered by the hypothetical AA-style mixture converges to the loss of the convex combination fast enough as β approaches 1.

Appendix D. Proof of Lemma 15

For every $\varepsilon > 0$ and $\gamma^* \in \Gamma$ the set $U(\gamma^*, \varepsilon) = \{\gamma \in \Gamma \mid \lambda(\omega, \gamma^*) < \lambda(\omega, \gamma) +$

ε for all $\omega \in \Omega\}$ is open. Indeed, λ is continuous and $U(\gamma^*, \varepsilon)$ is an intersection of finitely many inverse images of open sets.

For every finite $L > 0$ let $\Gamma_L = \{\gamma \in \Gamma \mid \lambda(\omega, \gamma) \le L \text{ for all } \omega \in \Omega\}$. Fix $\varepsilon > 0$. The union $\bigcup_{L>0} \bigcup_{\gamma^* \in \Gamma_L} U(\gamma^*, \varepsilon)$ is an open covering of Γ. Indeed, consider some $\gamma_0 \in \Gamma$. If the values $\lambda(\omega, \gamma_0)$ are finite for all ω, then γ_0 belongs to some Γ_L. If some of these values are infinite, γ_0 can still be approximated by predictions that can only lead to finite losses and therefore γ_0 belongs to $U(\gamma^*, \varepsilon)$ of some such γ^*.

Since Γ is compact, a finite subcovering exists and thus a finite L can be chosen. This proves the lemma.

Remark 20 *The lemma can also be proven by constructing a covering of the set of superprediction S. This way is slightly longer, but arguably more intuitive because the construction is done in $\mathbb{R}^M$.*

Let $|\Omega| = M$ and $\Omega = \{\omega^{(0)}, \omega^{(1)}, \dots, \omega^{(M-1)}\}$. Let Γ_L be as above and consider the sets $P_L = \left\{\left(\lambda(\omega^{(0)}, \gamma), \lambda(\omega^{(1)}, \gamma), \dots, \lambda(\omega^{(M-1)}, \gamma)\right) \mid \gamma \in \Gamma_L\right\}$.

For every $\varepsilon > 0$ let $V(L, \varepsilon)$ be the ε-vicinity of the set P_L, i.e., the union of all open balls of radius ε centred on points from P_L. Finally, let $S(L, \varepsilon) = \{X \in [-\infty, +\infty]^M \mid X \ge Y \text{ for some } Y \in V_{L,\varepsilon}\}$.

It is easy to check that for every $\epsilon > 0$ we have $S \subseteq \bigcup_{L>0} S(L, \varepsilon)$. One can show that this covering has a finite subcovering by considering the image under the transformation $\mathfrak{B}_\beta$ (see Subsection 2.4) with some $\beta \in (0, 1)$.

Appendix E. Choosing the Sequences

Take $M_0 = N_1$ and $M_j = N_{j+1} - N_j$, $j = 1, 2 \dots$ Let a positive integer n be such that $N_k < n \le N_{k+1}$ (see Figure 9). Applying (8) yields

$$\mathrm{Loss}^{\mathfrak{G}}_{\mathfrak{M}}(n) \le \mathrm{Loss}_{\mathcal{E}^{(i)}}(n) + r(n)$$

for all $i = 1, 2, \dots, N$, where N is the number of experts and

$$r(n) = \sum_{j=0}^{k-1} M_j \varepsilon_j + \sum_{j=0}^{k-1} C_{\varepsilon_j} \sqrt{M_j} + \varepsilon_k (n - N_k) + C_{\varepsilon_k} \sqrt{n - N_k} \tag{21}$$

is the 'remainder' (we recall that $C_\varepsilon = 2L_\varepsilon^2 \sqrt{\ln N}$). Note that the former two terms correspond to the previous invocations of WAA and the later two correspond to the current invocation.

We shall formulate conditions sufficient for the terms in (21) to be of $o(n)$ order of magnitude. First note that

(1) $\lim_{j\to+\infty} \varepsilon_j = 0$

and $k = k(n) \to \infty$ as $n \to \infty$ is sufficient to ensure that $\varepsilon_k(n - N_k) = o(n)$ as $n \to \infty$. Secondly, if, moreover,

(2′) $\sum_{j=0}^{\infty} M_j = +\infty$

then $\sum_{j=0}^{k-1} M_j\varepsilon_j = o(n)$ by the following simple lemma.

Lemma 21 *If the series $\sum_{i=1}^{\infty} M_i$ diverges and $\alpha_i \to 0$, where all M_i and α_i are non-negative, then $\sum_{i=1}^{k} M_i\alpha_i = o\left(\sum_{i=1}^{k} M_i\right)$ as $k \to \infty$.*

PROOF of Lemma 21 Take a small $\varepsilon > 0$. There is positive integer l such that $\alpha_i < \varepsilon/2$ for all $i \geq l$. We thus have

$$\sum_{i=1}^{k} M_i\alpha_i \leq \sum_{i=1}^{l} M_i\alpha_i + \frac{\varepsilon}{2}\sum_{i=l}^{k} M_i$$

for all $k \geq l$. Since the series diverges, $\sum_{i=l}^{k} M_i$ tend to $+\infty$ as $k \to \infty$ and thus for sufficiently large k

$$\sum_{i=1}^{l} M_i\alpha_i \leq \frac{\varepsilon}{2}\sum_{i=l}^{k} M_i$$

and therefore

$$\sum_{i=1}^{k} M_i\alpha_i \leq \varepsilon\sum_{i=1}^{k} M_i \ .$$

□

Thirdly, the lemma implies that if, moreover,

(3) $C_{\varepsilon_j} \leq \sqrt[8]{M_j}$, $j = 0, 1, 2, \ldots$,

then $\sum_{j=0}^{k-1} C_{\varepsilon_j}\sqrt{M_j} \leq \sum_{j=0}^{k-1} M_j/M_j^{3/8} = o(n)$.

It remains to consider the last term in (21). There are two cases, either $n - N_k \leq M_k^{3/4}$ or $n - N_k > M_k^{3/4}$. In the former case we get

$$\frac{1}{n}C_{\varepsilon_k}\sqrt{n - N_k} \leq \frac{M_k^{1/8}\sqrt{n - N_k}}{N_k} \leq \frac{M_k^{1/8}M_k^{3/8}}{M_{k-1}} = \frac{\sqrt{M_k}}{M_{k-1}} \ ,$$

while in the latter case we get

$$\frac{1}{n} C_{\varepsilon_k} \sqrt{n - N_k} \le \frac{M_k^{1/8} \sqrt{M_k}}{M_k^{3/4}} = \frac{1}{M_k^{1/8}} \ .$$

To ensure the convergence to 0, it is sufficient to add

(4) $M_{j-1} \ge M_j^{3/4}$, $j = 1, 2, \ldots$

and to replace $(2')$ with a stronger requirement

(2) $M_j \to +\infty$, $j \to \infty$.

Let us show that conditions (1)–(4) are compatible, i.e., construct the sequences ε_j and M_j. Let $M_0 = \max(2, \lceil C_{\varepsilon_0}^8 \rceil)$ and $M_{j+1} = \lfloor M_j^{4/3} \rfloor$, $j = 0, 1, 2, \ldots$ The sequence ε_j is constructed as follows. Suppose that all ε_j have been constructed for $j \le k$. If $C_{\varepsilon_k/2} \le M_k^{1/8}$, we let $\varepsilon_{k+1} = \varepsilon_k/2$; otherwise we let $\varepsilon_{k+1} = \varepsilon_k$. Since $M_k \to +\infty$ and C_ε is finite for every $\varepsilon > 0$, we shall be able to divide ε_k by 2 eventually and thus ensure that $\varepsilon_j \to 0$ as $j \to +\infty$.

Acknowledgements

We would like to thank participants of the Kolmogorov seminar on complexity theory at the Moscow State University and Alexander Shen in particular for useful suggestions that allowed us to simplify the WAA. We would also like to thank Volodya Vovk for suggesting an idea that helped us to strengthen an upper bound on the performance of WAA.

We are grateful to anonymous COLT and JCSS referees for their detailed comments. The paper has been greatly improved due to referees' suggestions.

References

[1] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, M. K. Warmuth, How to use expert advice, Journal of the ACM 44 (3) (1997) 427–485.

[2] D. Haussler, J. Kivinen, M. K. Warmuth, Sequential prediction of individual sequences under general loss functions, IEEE Transactions on Information Theory 44 (5) (1998) 1906–1925.

[3] N. Cesa-Bianchi, G. Lugosi, Prediction, Learning, and Games, Cambridge University Press, 2006.

[4] V. Vovk, Aggregating strategies, in: Proceedings of the 3rd Annual Workshop on Computational Learning Theory, Morgan Kaufmann, San Mateo, CA, 1990, pp. 371–383.

[5] V. Vovk, A game of prediction with expert advice, Journal of Computer and System Sciences 56 (1998) 153–173.

[6] M. Hutter, J. Poland, Adaptive online prediction by following the perturbed leader, Journal of Machine Learning Research 6(Apr) (2005) 639–660.

[7] Y. Kalnishkan, M. V. Vyugin, On the absence of predictive complexity for some games, in: Algorithmic Learning Theory, 13th International Conference, Proceedings, Vol. 2533 of Lecture Notes in Artificial Intelligence, Springer, 2002, pp. 164–172.

[8] D. Blackwell, M. A. Girshik, Theory of Games and Statistical Decisions, Wiley, 1954.

[9] H. G. Eggleston, Convexity, Cambridge University Press, Cambridge, 1958.

[10] Y. Kalnishkan, M. V. Vyugin, Mixability and the existence of weak complexities, in: Computational Learning Theory, 15th Annual Conference, Proceedings, Vol. 2375 of Lecture Notes in Artificial Intelligence, Springer, 2002, pp. 105–120.